UNIVERZITA PALACKÉHO V OLOMOUCI
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

# BAKALÁŘSKÁ PRÁCE

## Mnohorozměrné normální rozdělení a jeho aplikace

Vedoucí bakalářské práce:
**RNDr. Karel Hron, Ph.D.**
Rok odevzdání: 2009

Vypracovala:
**Michaela Šuková**
ME, III. ročník

**Prohlášení**

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana RNDr. Karla Hrona, Ph.D. a s použitím uvedené literatury.

V Olomouci dne 16. dubna 2009

**Poděkování**

Na tomto místě bych chtěla poděkovat především svému vedoucímu diplomové práce panu RNDr. Karlu Hronovi, Ph.D., že měl se mnou dostatek trpělivosti, aby mi pomohl dovézt tuto práci ke zdárnému konci. Také bych ráda poděkovala své rodině a přátelům, že mě po celou dobu studia podporovali.

# Obsah

# Úvod

Jako téma své bakalářské práce mě nejvíce zaujala možnost zabývat se normálním rozdělením, se kterým jsem se více či méně setkávala po celou dobu studia, a možností jeho aplikace pro speciální typ mnohorozměrných dat, tzv. kompoziční data, což je oblast statistiky vyvinutá v 80. letech minulého století. Svoji bakalářskou práci jsem rozdělila do dvou kapitol, popisujících mnohorozměrné normální rozdělení a kompoziční data. První kapitola se skládá ze tří částí. V první z nich se zabývám momentovou generující funkcí, její definicí a vlastnostmi pro diskrétní i spojité náhodné veličiny včetně názorných výpočtů. Ve druhé části této kapitoly je popsáno normální rozdělení za použití momentové generující funkce pro jeho odvození. V poslední části této kapitoly se věnuji samotnému mnohorozměrnému normálnímu rozdělení, opět s užitím vlastností momentové generující funkce. Druhá kapitola se zabývá prací s kompozičními daty, náhodnými vektory nesoucími pouze relativní informaci, a možnému zavedení normálního rozdělení v tomto speciálním případě. Přitom se též blíže zmiňuji o simplexu jako výběrovém prostoru kompozičních dat a celou situaci jsem opět ilustrovala na názorných příkladech.

# 1. Mnohorozměrné normální rozdělení

## 1.1. Momentová generující funkce

V základním kurzu teorie pravděpodobnosti jsme se seznámili s distribuční funkcí a pravděpodobnostní funkcí, resp. hustotou, jako variantami popisu rozdělení pravděpodobnosti diskrétních a spojitých náhodných veličin. Další možností je též momentová generující funkce, kterou si představíme v následující definici [6].

**Definice 1** (momentová generující funkce) Nechť $X$ je náhodná veličina taková, že pro nějaké $h > 0$ existuje střední hodnota náhodné veličiny $e^{tX}$, kde $-h < t < h$. Momentová generující funkce náhodné veličiny $X$ je definována jako reálná funkce $M(t) = \mathsf{E}(e^{tX})$, pro $-h < t < h$.

Pro další úvahy je potřeba, aby byla momentová generující funkce definována na otevřeném okolí kolem nuly, které bude zahrnovat interval $(-h, h)$ pro nějaké $h > 0$. Také je zřejmé, že pokud položíme $t = 0$, potom se momentová funkce $M(0)$ bude rovnat jedné, tzn. $M(0) = 1$. Musíme však ještě jednou poznamenat, že předpokladem je právě její výskyt v otevřeném intervalu kolem nuly. Z tohoto důvodu je zřejmé, že ne každé rozdělení má momentovou generující funkci. Jestliže budeme hovořit o více náhodných veličinách, pak budeme pro $M$ používat označení $M_X$ jako momentová generující funkce náhodné veličiny $X$.

Nechť $X$ a $Y$ jsou náhodné veličiny s momentovými generujícími funkcemi. Jestliže mají stejné rozdělení pravděpodobností, tj. $F_X(z) = F_Y(z)$, pro všechna $z$ reálná, potom v okolí nuly zřejmě platí $M_X(t) = M_Y(t)$. Avšak jednou z nejdůležitějších vlastností momentové generující funkce je, že opak

výroku uvedeného výše je také pravdivý, tedy momentové generující funkce popisují jednoznačně rozdělení náhodných veličin. Tento poznatek shrneme v následující větě [6]:

**Věta 1** *Nechť $X$ a $Y$ jsou náhodné veličiny s momentovými generujícími funkcemi $M_X$ a $M_Y$, které existují v otevřeném intervalu kolem nuly. Pak $F_X(z) = F_Y(z)$ pro každé $z \in \boldsymbol{R}$ tehdy, a jen tehdy, pokud platí $M_X(t) = M_Y(t)$, pro všechna $t \in (-h, h)$, kde $h > 0$.*

Ověříme tvrzení věty 1 na následujícím příkladě.

**Příklad 1** *Nechť*

$$M(t) = \frac{1}{10}e^t + \frac{2}{10}e^{2t} + \frac{3}{10}e^{3t} + \frac{4}{10}e^{4t}$$

*je pro všechna $t \in \boldsymbol{R}$ momentová generující funkce náhodné veličiny $X$ diskrétního typu. Jestliže $p(x)$ je její pravděpodobnostní funkce, pak*

$$M(t) = \sum_x e^{tx} p(x),$$

*tedy pro realizace $a_1, a_2, a_3, ...$ náhodné veličiny $X$ dostaneme*

$$\frac{1}{10}e^t + \frac{2}{10}e^{2t} + \frac{3}{10}e^{3t} + \frac{4}{10}e^{4t} = p(a_1)e^{a_1 t} + p(a_2)e^{a_2 t} + ...$$

*Protože rovnost platí pro všechna reálná $t$, pravá strana musí být složena ze čtyř výrazů a každý z nich roven odpovídajícímu na levé straně. Tedy pro $a_1 = 1$, dostaneme $p(a_1) = \frac{1}{10}$; analogicky pak $a_2 = 2$, $p(a_2) = \frac{2}{10}$; $a_3 = 3$,*

$p(a_3) = \frac{3}{10}$; $a_4 = 4$, $p(a_4) = \frac{4}{10}$. *Zjednodušeně řečeno, pravděpodobnostní funkce náhodné veličiny* $X$ *je rovna*

$$p(x) = \begin{cases} \frac{x}{10}, & x = 1, 2, 3, 4, \\ 0, & \text{jinak.} \end{cases}$$

*Dále předpokládejme náhodnou veličinu* $X$ *spojitého typu a nechť*

$$M(t) = \frac{1}{1-t}, \ t < 1,$$

*je momentová generující funkce náhodné veličiny* $X$. *Následkem toho dostaneme*

$$\frac{1}{1-t} = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx, \ t < 1.$$

*Odsud není na první pohled zřejmé, jak najdeme hustotu* $f(x)$. *Lze ovšem ověřit, že rozdělení s hustotu*

$$f(x) = \begin{cases} e^{-x}, & 0 < x < \infty \\ 0, & \text{jinak} \end{cases}$$

*má právě momentovou generující funkci* $M(t) = (1-t)^{-1}$ *pro* $t < 1$. *Náhodná veličina* $X$ *má tedy rozdělení s touto hustotou ve shodě s tvrzením Věty 1 o jednoznačnosti momentové generující funkce.*

Vzhledem k tomu, že rozdělení mající momentovou generující funkci $M(t)$ je touto funkcí plně určeno, není překvapením, že přímo z $M(t)$ získáme některé vlastnosti tohoto rozdělení. Například existence $M(t)$ pro $-h < t < h$ implikuje, že pro $t = 0$ existují derivace všech řádů $M(t)$. S využitím teorie diferenciálního a integrálního počtu by bylo možné ověřit, že lze zaměnit pořadí derivování a integrování, resp. diferencování a sumace v diskrétním případě.

Potom pro náhodnou veličinu $X$, která má spojité rozdělení pravděpodobnosti, platí

$$M'(t) = \frac{dM(t)}{dt} = \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f(x)\, dx = \int_{-\infty}^{\infty} x e^{tx} f(x)\, dx,$$

případně, jestliže $X$ je diskrétní náhodná veličina,

$$M'(t) = \frac{dM(t)}{dt} = \sum_{x} x e^{tx} p(x).$$

Položíme $t = 0$, můžeme tedy říct, že platí

$$M'(0) = \mathsf{E}(X) = \mu.$$

Druhá derivace $M(t)$ pak analogicky bude

$$M''(t) = \int_{-\infty}^{\infty} x^2 e^{tx} f(x)\, dx \quad \text{nebo} \quad M''(t) = \sum_{x} x^2 e^{tx} p(x),$$

tedy

$$M''(0) = \mathsf{E}(X^2).$$

Pro rozptyl $\mathrm{var}(X) = \sigma^2$ veličiny $X$ tedy platí

$$\sigma^2 = \mathsf{E}(X^2) - \mu^2 = M''(0) - [M'(0)]^2.$$

Například, jestliže $M(t) = (1-t)^{-1},\ t < 1$ z příkladu uvedeného výše, pak platí

$$M'(t) = (1-t)^{-2} \text{ a } M''(t) = 2(1-t)^{-3}.$$

Odtud přímo plyne

$$\mu = M'(0) = 1 \quad \text{a} \quad \sigma^2 = M''(0) - \mu^2 = 2 - 1 = 1.$$

Samozřejmě, $\mu$ a $\sigma^2$ bychom mohli též spočítat přímo z hustoty jako

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \quad \text{a} \quad \sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2.$$

Někdy bývá jeden postup snadnější než druhý. Obecně, nechť $m$ je kladné celé číslo a označme $M^{(m)}(t)$ $m$-tou derivaci momentové generující funkce $M(t)$. Pak opakovanou derivací podle $t$ dostaneme

$$M^{(m)}(0) = \mathsf{E}(X^m).$$

Poznamenejme, že integrály (nebo také součty) typu

$$\mathsf{E}(X^m) = \int_{-\infty}^{\infty} x^m f(x)\, dx \text{ nebo } \mathsf{E}(X^m) = \sum_x x^m p(x)$$

jsou v mechanice nazývány momenty. Vzhledem k tomu že $M(t)$ generuje hodnoty $\mathsf{E}(X^m)$, kde $m = 1, 2, 3, \ldots$, nazývá se momentovou generující funkcí. Potom tedy opravdu můžeme nazvat číslo $\mathsf{E}(X^m)$, které odpovídá $m$-té derivaci momentové generující funkce, $m$-tým (obecným) momentem rozdělení nebo $m$-tým momentem náhodné veličiny $X$.

Momentová generující funkce sdruženého rozdělení $n$ náhodných veličin $X_1, X_2, \ldots, X_n$ je definována jako

$$\mathsf{E}[\exp(t_1 X_1 + t_2 X_2 + \ldots + t_n X_n)],$$

existuje–li pro $-h_i < t_i < h_i, i = 1, 2, \ldots, n$, kde $h_i$ je kladné pro každé $i$. Tuto střední hodnotu, kterou označíme jako $M(t_1, t_2, \ldots, t_n)$, nazýváme momentovou generující funkcí sdruženého rozdělení náhodných veličin $X_1, X_2, \ldots, X_n$.

Například, momentová generující funkce marginálního rozdělení náhodné veličiny $X_i$ je z tohoto pohledu rovna $M(0,\ldots,0,t_i,0,\ldots,0)$ pro $i=1,2,\ldots,n$, dále marginálnímu rozdělení veličin $X_i$ a $X_j$, $1 \leq i,j \leq n, i \neq j$, odpovídá $M(0,\ldots,0,t_i,0,\ldots,0,t_j,0,\ldots,0)$, atd., a proto existence rozkladu

$$M\left(t_1,t_2,\ldots,t_n\right) = \prod_{i=1}^{n} M\left(0,\ldots,0,t_i,0,\ldots,0\right)$$

je nutnou a také postačující podmínkou pro nezávislost náhodných veličin $X_1,X_2,\ldots,X_n$. Poznamenejme, že sdruženou momentovou generující funkci lze zapsat i vektorově jako

$$M\left(\mathbf{t}\right) = \mathsf{E}\left[\exp\left(\mathbf{t}'\mathbf{X}\right)\right]$$

pro $\mathbf{X} = (X_1,X_2,\ldots,X_n)'$, $\mathbf{t} = (t_1,t_2,\ldots,t_n)'$ a $\mathbf{t} \in B \subset \boldsymbol{R}^n$, kde

$$B = \left\{\mathbf{t} : -h_i < t_i < h_i, i = 1,2,\ldots,n\right\}.$$

## 1.2. Jednorozměrné normální rozdělení

Normální rozdělení pravděpodobnosti představuje důležitou třídu rozdělení pro statistické usuzování včetně mnoha aplikací, například je též odrazovým můstkem pro velmi elegantní a explicitní teorii lineárních modelů. V dalším textu nejdříve zavedeme tzv. normované normální rozdělení a skrze něj potom obecné normální rozdělení [6]. Uvažujme tedy integrál

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz.$$

Tento integrál existuje, protože argument je kladná spojitá funkce, která je omezená integrovatelnou funkcí, konkrétně

$$0 < \exp\left(\frac{-z^2}{2}\right) < \exp\left(-|z|+1\right), \quad -\infty < z < \infty,$$

a přitom

$$\int_{-\infty}^{\infty} \exp\left(-|z|+1\right) dz = 2e.$$

Pro určení hodnoty $I$ poznamenejme, že $I > 0$ a $I^2$ může být zapsána jako

$$I^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(\frac{-z^2+w^2}{2}\right) dz\, dw.$$

Tento dvojný integrál lze vyřešit převedením do polárních souřadnic.
Položíme–li

$$z = r \cdot \cos\theta \quad \text{a} \quad w = r \cdot \sin\theta, \quad r > 0, \quad \theta \in \langle 0, 2\pi \rangle,$$

dostaneme

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} \cdot r\, dr\, d\theta = \frac{1}{2\pi} \int_0^{2\pi} d\theta = 1.$$

Protože integrand u $I$ je kladný pro všechna reálná čísla a $I = 1$, je hustotou nějaké spojité náhodné veličiny, kterou označíme $Z$. Dohromady, $Z$ má hustotu

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right), -\infty < z < \infty. \tag{1}$$

Momentová generující funkce náhodné veličiny $Z$ je rovna

$$\mathsf{E}\left[\exp(tZ)\right] = \int_{-\infty}^{\infty} \exp(tz) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz =$$

$$= \exp\left(\frac{1}{2}t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}(z-t)^2\right) dz =$$

$$= \exp\left(\frac{1}{2}t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) dw,$$

přičemž jsme v posledním integrálu provedli substituci $w = z - t$. Hodnota posledně uvedeného integrálu je rovna jedné, a tedy pro $Z$ platí

$$M_Z(t) = \exp\left(\frac{1}{2}t^2\right),$$

kde $-\infty < t < \infty$. Vypočteme první a druhou derivaci,

$$M'_Z(t) = t \cdot \exp\left(\frac{1}{2}t^2\right)$$

a

$$M''_Z(t) = \exp\left(\frac{1}{2}t^2\right) + t^2 \cdot \exp\left(\frac{1}{2}t^2\right).$$

Položením $t = 0$ tak určíme střední hodnotu a rozptyl náhodné veličiny $Z$,

$$\mathsf{E}(Z) = 0 \quad \text{a} \quad \text{var}(Z) = 1. \tag{2}$$

Dále definujeme spojitou náhodnou veličinu $X$ jako

$$X = bZ + a, \quad \text{kde } b > 0\ .$$

Přitom $X$ je lineární transformací náhodné veličiny $Z$, tedy hustota bude

ve tvaru

$$f_X(x) = \frac{1}{\sqrt{2\pi} \cdot b} \exp\left(-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right)$$

pro $-\infty < x < \infty$. Užitím (2) dostaneme $\mathsf{E}(X) = a$ a $\mathrm{var}(X) = b^2$. Následně nahradíme ve výrazu pro hustotu $a$ parametrem $\mu$ a $b$ kladným parametrem $\sigma^2$. Obdržíme tak následující definici [8]:

**Definice 2** Řekneme, že náhodná veličina $X$ má normální rozdělení s parametry $\mu$ a $\sigma^2$, jestliže pro její hustotu platí

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad \text{pro} - \infty < x < \infty.$$

Parametry $\mu$ a $\sigma^2$ tedy představují střední hodnotu a rozptyl náhodné veličiny $X$. V dalším textu budeme často zapisovat, že $X$ má rozdělení $N(\mu, \sigma^2)$. V tomto označení má náhodná veličina $Z$ rozdělení $N(0,1)$ a nazýváme je *normované normální rozdělení.*

Pro určení momentové generující funkce náhodné veličiny $X$ užijeme vztahu $X = \sigma Z + \mu$ a momentové generující funkce pro $Z$, tedy

$$M_X(t) = \mathsf{E}\left[\exp(tX)\right] = \mathsf{E}\left[\exp(t(\sigma Z + \mu))\right] =$$

$$= \exp(\mu t)\mathsf{E}\left[\exp(t\sigma Z)\right] = \exp(\mu t)\exp\left(\frac{1}{2}\sigma^2 t^2\right) =$$

$$= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right),$$

kde $-\infty < t < \infty$. Předchozí úvahy tedy vedou k závěru, že $X$ má rozdělení $N(\mu, \sigma^2)$ tehdy, a jen tehdy, má–li $Z = \frac{X-\mu}{\sigma}$ rozdělení $N(0,1)$.

Normálnímu rozdělení se také říká rozdělení chyb nebo se nazývá Gaussovo rozdělení podle Karla Friedricha Gausse, který jej v roce 1809 zavedl. Tak jsme se dostali ke známému termínu Gaussova křivka, který se užívá pro označení hustoty. Graf hustoty normálního rozdělení s charakteristickým zvonovitým tvarem má jediný vrchol v $\mu$ a je symetrický kolem střední hodnoty, která je zároveň rovna modu a mediánu (viz Obrázek 1 pro případ normovaného normálního rozdělení) [7]. Připomeňme, že modus u spojitého rozdělení je právě bod, ve kterém má hustota $f_X$ lokální maximum. Je zřejmé, že platí rovnost

$$f_X(\mu - x) = f_X(\mu + x), \forall x \in \boldsymbol{R}.$$

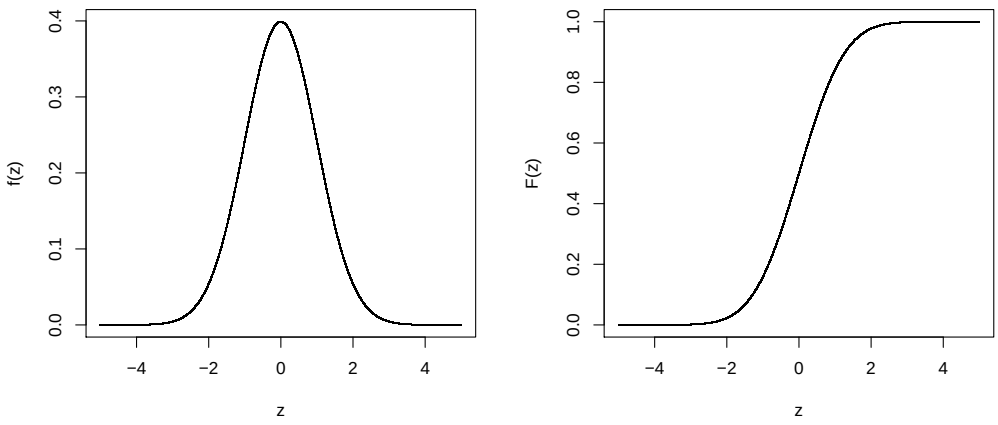
Oproti tomu rozptyl nám ukazuje, zda hodnoty náhodné veličiny se více či méně koncentrují kolem střední hodnoty $\mu$. Změna parametru $\mu$ potom znamená posun $f_X$ ve směru osy $x$, oproti tomu $\sigma$ určuje míru 'zploštělosti' křivky.
Provedením substituce $Z = \frac{X-\mu}{\sigma}$ lze distribuční funkci $X$ vyjádřit jako

$$F_X(x) = \mathsf{P}(X \leq x) = \mathsf{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

kde

$$\Phi(z) = \mathsf{P}(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{u^2}{2}} du$$

je distribuční funkce náhodné veličiny $Z$ (viz Obrázek 1). Této transfromace se často využívá, potřebujeme–li hodnotu $F_X(x)$ najít právě ve statistických tabulkách. Navíc využití vlastnosti symetrie hustoty normovaného normálního rozdělení okolo nuly nám umožní vypočítat hodnotu distribuční




Obrázek 1: Hustota a distribuční funkce normálního normovaného rozdělení

funkce $\Phi(-z)$, kde $z > 0$, jako

$$\Phi\left(-z\right) = 1 - \Phi\left(z\right).$$

Distribuční funkce normálního rozdělení náhodné veličiny $X$ je tak dána předpisem

$$F_X\left(x\right) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du,$$

bohužel ji tedy není možné vyjádřit pomocí elementárních funkcí a její hodnoty hledáme v tabulkách či pomocí statistického softwaru (viz předchozí strana). Použití distribuční funkce normovaného normálního rozdělení pro výpočet pravděpodobností, s kterými náhodná veličina $X$ nabývá svých realizací, si ukážeme na následujících příkladech.

**Příklad 2** *Nechť* $X \sim N(2; 25)$*, pak*

$$\mathsf{P}\left(0 < X < 10\right) = \Phi\left(\frac{10-2}{5}\right) - \Phi\left(\frac{0-2}{5}\right) =$$

$$= \Phi(1,6) - \Phi(-0,4) = 0,945 - (1 - 0,655) = 0,600.$$

**Příklad 3** *Nechť $X$ má normální rozdělení $N(\mu, \sigma^2)$, potom*

$$\mathsf{P}(\mu - 2\sigma < X < \mu + 2\sigma) = \Phi\left(\frac{\mu + 2\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 2\sigma - \mu}{\sigma}\right) =$$

$$= \Phi(2) - \Phi(-2) = 0.977 - (1 - 0.977) = 0.954.$$

Poznamenejme, že z normálního rozdělení je odvozena řada rozdělení dalších např. $\chi^2$ rozdělení, Studentovo nebo také Fisherovo rozdělení, které mají mimořádně důležitou roli v matematické statistice [8].

## 1.3. Mnohorozměrné normální rozdělení

V této kapitole se budeme zabývat již samotným mnohorozměrným normálním rozdělením. Zavedeme jej obecně pro $n$–rozměrný náhodný vektor, ale také speciálně zmíníme dvourozměrný případ, tzn. $n = 2$. Stejně jako v kapitole o jednorozměrném normálním rozdělení začneme nejprve normovaným rozdělením a poté přejdeme k obecnému případu [6]. Přitom budeme s výhodou užívat maticového zápisu.

Uvažujme náhodný vektor $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)'$, kde $Z_1, Z_2, ..., Z_n$ jsou nezávislé náhodné veličiny mající rozdělení $N(0,1)$. Pak hustota náhodného vektoru $\mathbf{Z}$ je

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} z_i^2\right) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot \exp\left(-\frac{1}{2} \sum_{i=1}^{n} z_i^2\right) =$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot \exp\left(-\frac{1}{2} \mathbf{z}'\mathbf{z}\right), \text{pro } \mathbf{z} \in \boldsymbol{R}^n.$$

Protože $Z_i$ mají střední hodnotu rovnu nule, rozptyl $\mathrm{var}(Z_i) = 1$ a jsou nekorelované, platí pro střední hodnotu a varianční matici náhodného vektoru $\mathbf{Z}$

$$\mathsf{E}(\mathbf{Z}) = \mathbf{0} \text{ a } \mathrm{var}(\mathbf{Z}) = \mathbf{I}_n,$$

kde $\mathbf{I}_n$ značí jednotkovou matici řádu $n$. Vzpomeňme, že momentové generující funkce veličin $Z_i$ jsou rovny

$$M_{Z_i}(t_i) = \exp\left(\frac{t_i^2}{2}\right), \quad i = 1, \ldots, n.$$

Tedy, protože $Z_i$ jsou nezávislé, momentová generující funkce náhodného vektoru $\mathbf{Z}$ je rovna

$$M_{\mathbf{Z}}(\mathbf{t}) = \mathsf{E}\left[\exp\left(\mathbf{t}'\mathbf{Z}\right)\right] = \mathsf{E}\left[\prod_{i=1}^{n} \exp\left(t_i Z_i\right)\right] =$$

$$= \prod_{i=1}^{n} \mathsf{E}\left[\exp(t_i Z_i)\right] = \exp\left(\frac{1}{2}\sum_{i=1}^{n} t_i^2\right) = \exp\left(\frac{1}{2}\mathbf{t}'\mathbf{t}\right)$$

pro všechna $\mathbf{t} \in \boldsymbol{R}^n$. Říkáme, že náhodný vektor $\mathbf{Z}$ má *mnohorozměrné normální rozdělení* se střední hodnotou rovnou nulovému vektoru $\mathbf{0}$ a varianční maticí $\mathbf{I}_n$, což zjednodušeně zapisujeme jako $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$.

V obecném případě předpokládejme, že $\boldsymbol{\Sigma}$ je symetrická pozitivně semidefinitní matice typu $n \times n$, , tj. $\forall\, \mathbf{x} \in \boldsymbol{R}^n, \mathbf{x}'\,\boldsymbol{\Sigma}\,\mathbf{x} \geq 0$. Potom, díky znalostem lineární algebry, můžeme $\boldsymbol{\Sigma}$ vždy rozložit jako součin

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma},$$

kde $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ je diagonální matice, $\lambda_1 \geq \lambda_2 \geq, ..., \geq \lambda_n \geq 0$ jsou nezáporná vlastní čísla matice $\boldsymbol{\Sigma}$ a sloupce matice $\boldsymbol{\Gamma}$, $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$,

jsou odpovídající vlastní vektory. Tento rozklad se nazývá *spektrální rozklad matice* $\mathbf{\Sigma}$. Matice $\mathbf{\Gamma}$ je ortogonální, tj. $\mathbf{\Gamma}^{-1} = \mathbf{\Gamma}'$, a tedy $\mathbf{\Gamma}\mathbf{\Gamma}' = \mathbf{I} = \mathbf{\Gamma}'\mathbf{\Gamma}$. Spektrální rozklad můžeme ovšem zapsat i jiným způsobem, a to jako

$$\mathbf{\Sigma} = \mathbf{\Gamma}'\mathbf{\Lambda}\mathbf{\Gamma} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i'.$$

Protože $\lambda_i$ jsou nezáporná, můžeme definovat diagonální matici

$$\mathbf{\Lambda}^{\frac{1}{2}} = \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_n}\right);$$

pak z ortogonality $\mathbf{\Gamma}$ plyne

$$\mathbf{\Sigma} = \mathbf{\Gamma}'\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Gamma}$$

a definujeme *odmocninu pozitivně semidefinitní matice* $\mathbf{\Sigma}$ jako

$$\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Gamma}'\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Gamma}.$$

Poznamenejme, že $\mathbf{\Sigma}^{\frac{1}{2}}$ je opět symetrická a pozitivně semidefinitní. Dále předpokládejme, že $\mathbf{\Sigma}$ je pozitivně definitní, tj. všechna její vlastní čísla jsou kladná.

Zaveďme náhodný vektor $\mathbf{X} = \boldsymbol{\mu} + \mathbf{\Sigma}^{1/2}\mathbf{Z}$, kde $\boldsymbol{\mu}$ je n–rozměrný náhodný vektor. Potom pro jeho momentovou generující funkci s využitím $M_{\mathbf{Z}}(\mathbf{t})$ zřejmě obdržíme [6]

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\mathbf{\Sigma}\mathbf{t}\right).$$

Provedená úvaha vede k následující definici:

**Definice 3** Řekneme, že $n$–rozměrný náhodný vektor $\mathbf{X}$ má mnohorozměrné normální rozdělení s parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ , jestliže jeho momentová generující funkce je rovna

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right)$$

pro všechna $\mathbf{t} \in \boldsymbol{R}^n$, kde $\boldsymbol{\Sigma}$ je symetrická pozitivně definitní matice a $\boldsymbol{\mu} \in \boldsymbol{R}^n$. Pro jeho hustotu platí

$$f(\mathbf{x}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Pro zjednodušení zápisu budeme psát $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

**Věta 2** *Předpokládejme, že* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Nechť* $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, *kde* $\mathbf{A}$ *je číselná matice typu* $m \times n$ *a* $\mathbf{b} \in \boldsymbol{R}^m$, *pak* $\mathbf{Y}$ *má normální rozdělení* $N_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

Tato věta říká, že lineární transformace normálně rozděleného náhodného vektoru má též normální rozdělení s parametry, které vzniknou lineární transformací výchozích parametrů $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$. S využitím této věty lze snadno odvodit marginální rozdělení k mnohorozměrnému normálnímu rozdělení náhodného vektoru $\mathbf{X}$ [2]. Nechť $\mathbf{X}_1$ je podvektorem $\mathbf{X}$ dimenze $m < n$. Bez újmy na obecnosti můžeme psát

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix},$$

kde $\mathbf{X}_2$ má dimenzi $p = n - m$. Analogicky rozdělíme střední hodnotu a varianční matici náhodného vektoru $\mathbf{X}$ jako

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

a

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix};$$

poznamenejme, že $\boldsymbol{\Sigma}_{11}$ je varianční matice vektoru $\mathbf{X}_1$ a $\boldsymbol{\Sigma}_{12}$ obsahuje všechny kovariance mezi složkami vektorů $\mathbf{X}_1$ a $\mathbf{X}_2$. Nyní definujme matici

$$\mathbf{A} = [\mathbf{I}_m, \mathbf{O}_{mp}],$$

kde $\mathbf{O}_{mp}$ je matice samých nul typu $m \times p$. Pak zřejmě $\mathbf{X}_1 = \mathbf{AX}$ a z předchozí věty plyne následující důsledek:

**Důsledek 1** *Předpokládejme, že náhodný vektor* $\mathbf{X}$ *má* $n$*–rozměrné normální rozdělení se střední hodnotou* $\boldsymbol{\mu}$ *a varianční maticí* $\boldsymbol{\Sigma}$*, které jsou rozděleny výše uvedeným způsobem. Pak* $\mathbf{X_1}$ *má rozdělení* $N_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$*.*

Tento užitečný výsledek říká, že jakékoliv marginální rozdělení náhodného vektoru $\mathbf{X}$ je také normální, a dále jeho střední hodnota a varianční matice jsou asociovány s příslušným marginálním vektorem (maticí).

**Příklad 4** *V tomto příkladu budeme zkoumat situaci, kdy* $n = 2$*, tedy náhodný vektor má pouze dvě složky, a rozdělení v tomto případě se tedy nazývá dvourozměrné. Budeme také užívat obvyklý zápis* $(X, Y)'$ *namísto* $(X_1, X_2)'$*. Potom tedy předpokládejme, že* $(X, Y)'$ *má rozdělelní* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, kde*

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

*a*

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{22} & \sigma_2^2 \end{pmatrix}.$$

*V tomto případě* $\mu_1$ *a* $\sigma_1^2$ *jsou střední hodnota a rozptyl náhodné veličiny* $X$*,* $\mu_2$ *a* $\sigma_2^2$ *jsou střední hodnota a rozptyl náhodné veličiny* $Y$ *a* $\sigma_{12}$ *je kovariance mezi náhodnými veličinami* $X$ *a* $Y$*. Vzpomeňme, že* $\sigma_{12} = \rho\sigma_1\sigma_2$*, kde* $\rho$ *je korelační koeficient mezi* $X$ *a* $Y$*. Substitucí* $\rho\sigma_1\sigma_2$ *pro* $\sigma_{12}$ *v* $\mathbf{\Sigma}$ *lze snadno odvodit, že determinant* $\mathbf{\Sigma}$ *je roven* $\sigma_1^2\sigma_2^2(1-\rho^2)$*, kde* $\rho^2 \leq 1$ *. Pro zbytek tohoto příkladu budeme však předpokládat, že* $\rho^2 < 1$*. V tomto případě je tedy* $\mathbf{\Sigma}$ *invertibilní (a také pozitivně definitivní). Dále, vzhledem k tomu, že* $\mathbf{\Sigma}$ *je matice typu* $2 \times 2$*, můžeme její inverzi snadno určit jako*

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}.$$

*Užitím tohoto vztahu může být hustota náhodného vektoru* $(X,Y)'$ *zapsána ve tvaru*

$$f\left(x,y\right) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{q}{2}},$$

*kde*

$$q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right],$$

$-\infty < x < \infty, -\infty < y < \infty$*. Připomeňme, že pro* $X$ *a* $Y$ *nezávislé je jejich korelační koeficient roven nule. Pokud obě mají normální rozdělení, pak podle Důsledku 1 má* $X$ *rozdělení* $N(\mu_1, {\sigma_1}^2)$ *a* $Y$ *rozdělení* $N(\mu_2, {\sigma_2}^2)$*. Z tvaru hustoty* $f(x,y)$ *je přitom zřejmé, že pro* $\rho = 0$ *je sdružená hustota rovna součinu marginálních hustot, tedy* $X$ *a* $Y$ *jsou nezávislé.*

V případě dvourozměrné normality je tedy nezávislost ekvivalentní $\rho = 0$. Obdobné tvrzení lze potom odvodit i pro případ mnohorozměrného normálního rozdělení.

**Věta 3** *Nechť je dán náhodný vektor* $\mathbf{X}$*, který má normální rozdělení se střední hodnotou* $\boldsymbol{\mu}$ *a varianční maticí* $\boldsymbol{\Sigma}$ *a nechť* $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$*. Pak* $\mathbf{X}_1$ *a* $\mathbf{X}_2$ *jsou nezávislé tehdy, a jen tehdy, pokud* $\boldsymbol{\Sigma}_{12} = \mathbf{O}$*.*

# 2. Kompoziční data

## 2.1. Definice a základní vlastnosti

V následující kapitole se seznámíme se speciálním typem mnohorozměrných dat, tzv. kompozičními daty [1], která indukují odlišný výběrový prostor, simplex, na rozdíl od standardních náhodných vektorů. Nejprve přitom budeme uvažovat pouze nenáhodná kompoziční data a jako náhodné je zavedeme teprve v kapitole o normálním rozdělení na simplexu.

**Definice 4** [9] Sloupcový vektor $\mathbf{x} = (x_1, x_2, ..., x_D)'$ se nazývá $D$–složková kompozice, jestliže všechny jeho složky jsou kladná reálná čísla nesoucí pouze relativní informaci.

To znamená, že jak daná kompozice $\mathbf{x}$, tak i její kladný násobek nesou tutéž informaci. Tím se liší od běžných mnohorozměrných dat, které nesou informaci absolutní. Jako důsledek můžeme součet složek kompozic položit roven dané kladné konstantě $k$, která by v případě interpretace složek jako procentuálních podílů odpovídala $k = 1$ nebo $k = 100$.

Výběrový prostor $D$–složkových kompozic je tedy simplex

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, x_2, ..., x_D)', x_i > 0,\ i = 1, 2, ..., D;\ \sum_{i=1}^{D} x_i = k \right\}.$$

V případě trojsložkových kompozic tvoří simplex rovnostranný trojúhelník o výšce $k$ s vrcholy $A = [k, 0, 0]$, $B = [0, k, 0]$, $C = [0, 0, k]$. Pro složky kompozice $\mathbf{x} = (x_1, x_2, x_3)'$ pak platí, že $x_1$ je vzdálenost od strany $a$, $x_2$ je vzdálenost od strany $b$ a $x_3$ je vzdálenost od strany $c$. Pokud tento rovnostranný trojúhelník zobrazíme v rovině, vzniká graf, který nazýváme ternární diagram.

Simplex jako výběrový prostor kompozic indukuje přirozeně odlišnou geometrickou strukturu, to znamená, že je třeba zavést operace analogické sčítání vektorů a násobení skalárem v reálném prostoru, a dále skalární součin, normu a vzdálenost dvou kompozic. V tomto smyslu pak hovoříme o tzv. Aitchisonově geometrii na simplexu, kterou si nyní stručně představíme. Následující definice uzávěru je přirozeným důsledkem definice kompozičních dat [9]:

**Definice 5** Uzávěr kompozice $\mathbf{x}= (x_1, x_2, ..., x_D)'$, $x_i > 0, i = 1, 2, ..., D$, je definován jako

$$\mathcal{C}(\mathbf{x}) = \left( \frac{k \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{k \cdot x_2}{\sum_{i=1}^{D} x_i}, ..., \frac{k \cdot x_D}{\sum_{i=1}^{D} x_i} \right)'.$$

**Definice 6** Permutace kompozice $\mathbf{x} \in \mathcal{S}^D$ kompozicí $\mathbf{y} \in \mathcal{S}^D$ je dána vztahem

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left(x_1 y_1, x_2 y_2, ..., x_D y_D\right).$$

**Definice 7** Mocninnou transformaci kompozice $\mathbf{x} \in \mathcal{S}^D$ konstantou $\alpha \in \boldsymbol{R}$ definujeme jako

$$\alpha \odot \mathbf{x} = \mathcal{C}\left(x_1^\alpha, x_2^\alpha, ..., x_D^\alpha\right).$$

Simplex s operacemi permutace a mocninná transformace je vektorovým prostorem, tedy tyto operace mají analogické vlastnosti jako v případě standardních vektorů (komutativitu, asociativitu, distributivitu).

**Definice 8** Skalární součin $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ je definován jako

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

**Definice 9** Normu kompozice $\mathbf{x} \in \mathcal{S}^D$ definujeme

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j}\right)^2} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}.$$

**Definice 10** Vzdálenost mezi $\mathbf{x}$ a $\mathbf{y} \in \mathcal{S}^D$ je dána

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\right)^2} = \|\mathbf{x} - \mathbf{y}\|_a.$$

Normu a skalární součin z výše uvedených definic nazýváme někdy Aitchisonovou normou a Aitchisonovým skalárním součinem podle zakladatele moderního, tzv. logratio přístupu ke statistické analýze kompozičních dat, a umožňují nám zavést pojem ortonormální báze na simplexu, který má dimenzi $D-1$. Tu tvoří takové kompozice $\mathbf{u}_1, ..., \mathbf{u}_{D-1}$, pro které platí $\langle \mathbf{u}_i, \mathbf{u}_j \rangle_a = 0$ pro $i \neq j$ a $\|\mathbf{u}_i\|_a = 1$ pro $i, j = 1, ..., D-1$. Pomocí ortonormální báze na simplexu pak zavádíme tzv. isometric logratio ($ilr$) transformaci [3], která

kompozici $\mathbf{x} \in \mathcal{S}^D$ zobrazí do $D-1$ rozměrného reálného prostoru $\boldsymbol{R}^{D-1}$ a lze ji vyjádřit jako

$$ilr\left(\mathbf{x}\right) = \left(\left\langle \mathbf{x}, \mathbf{u}_1 \right\rangle_a, \left\langle \mathbf{x}, \mathbf{u}_2 \right\rangle_a, ..., \left\langle \mathbf{x}, \mathbf{u}_{D-1} \right\rangle_a\right) = \left(z_1, \ldots, z_{D-1}\right) = \mathbf{z}.$$

Složky vektoru $\mathbf{z}$ tedy tvoří souřadnice kompozice $\mathbf{x}$ vzhledem k bázi $\mathbf{u}_1, \ldots, \mathbf{u}_{D-1}$. Pro jednu konkrétní volbu báze $\mathbf{u}_1, \ldots, \mathbf{u}_{D-1}$ [5] můžeme tuto transformaci explicitně zapsat

$$z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \text{pro } i = 1, \ldots, D-1. \tag{3}$$

Výhoda *ilr* transformace je, že zobrazuje kompozice ze simplexu do reálného prostoru, kde již můžeme provádět standardní statistickou analýzu při zachování operací na simplexu, tedy

$$ilr\left(\mathbf{x} \oplus \mathbf{y}\right) = ilr\left(\mathbf{x}\right) + ilr\left(\mathbf{y}\right), ilr\left(\alpha \odot \mathbf{x}\right) = \alpha \cdot ilr\left(\mathbf{x}\right),$$

$$\left\langle \mathbf{x}, \mathbf{y} \right\rangle_a = \left\langle ilr\left(\mathbf{x}\right), ilr\left(\mathbf{y}\right) \right\rangle,$$

kde skalární součin vpravo je standardní euklidovský skalární součin. Místo *ilr* transformace též často hovoříme o vyjádření kompozice $\mathbf{x}$ v souřadnicích. Označíme–li $\mathbf{z} = ilr(\mathbf{x})$, inverzní zobrazení je potom $\mathbf{x} = ilr^{-1}(\mathbf{z})$, opět při konkrétní volbě báze výše jako $\mathbf{x} = \mathcal{C}\left(x_1, \ldots, x_D\right)$, kde

$$x_i = \exp\left(\sum_{j=1}^{D} \frac{z_j}{\sqrt{j\left(j+1\right)}} - \sqrt{\frac{i-1}{i}} z_{i-1}\right), \text{ pro } z_0 = z_D = 0,\ i = 1, \ldots, D.$$

Možnostem statistické analýzy *ilr* transformovaných kompozic se věnuje řada publikací. My se ovšem zaměříme pouze na zavedení normálního rozdělení na simplexu.

## 2.2. Normální rozdělení na simplexu

Také pro kompoziční data hraje normální rozdělení náhodných kompozic zcela zásadní roli pro možnosti statistické analýzy v souřadnicích. Tomu je též přizpůsobena následující definice [9]:

**Definice 11** Nechť je dán náhodný vektor $\mathbf{X}$, jehož výběrovým prostorem je simplex (tj. náhodná kompozice), pak říkáme, že $\mathbf{X}$ má normální rozdělení na $\mathcal{S}^D$ tehdy, a jen tehdy, když náhodný vektor ortonormálních souřadnic $\mathbf{X}^* = ilr(\mathbf{X})$ má mnohorozměrné normální rozdělení na $\boldsymbol{R}^{D-1}$.

Abychom mohli charakterizovat normální rozdělení, potřebujeme znát jeho parametry, čemuž ve standardním případě odpovídají střední hodnota $\boldsymbol{\mu}$ a varianční matice $\boldsymbol{\Sigma}$ [4]. Poznamenejme přitom, že volba souřadnic (tedy ortonormální báze na simplexu) nemá na existenci normálního rozdělení vliv, protože tato znamená pouze ortonormální rotaci transformovaných dat, a tedy opět normální rozdělení, pouze s jinými parametry. Pro ortogonální rotaci $\mathbf{Y}^*$ kompozice $\mathbf{X}$, vyjádřené v souřadnicích, tedy $\mathbf{X}^*$, totiž platí $\mathbf{Y}^* = \mathbf{P}\mathbf{X}^*$, kde $\mathbf{P}\mathbf{P}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$, a tedy dle Věty 2 ověříme normální rozdělení s parametry $\mathbf{P}\boldsymbol{\mu}$ a $\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}'$. Podobu normálního rozdělení se zvolenými parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ v souřadnicích i na simplexu si budeme ilustrovat na následujících příkladech. Na Obrázku 2 je zachycen výsledek simulace realizací normovaného normálního rozdělení kompozice $\mathbf{X}$ (tj. s $\boldsymbol{\mu}$ rovným nulovému vektoru a $\boldsymbol{\Sigma}$ jednotkové matici). Vliv parametrů je zřejmý z vyjádření kompozice $\mathbf{X}$ v souřadnicích, přičemž bylo využito ortonormální báze, odpovídající vztahům (3), tedy $z_1 = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$, $z_2 = \sqrt{\frac{2}{3}} \ln \frac{\sqrt{x_1 x_2}}{x_3}$. Na Obrázku 3 je zachycena změna situace pro parametry

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ a } \boldsymbol{\Sigma} = \begin{pmatrix} 4 & -1,9 \\ -1,9 & 1 \end{pmatrix}$$

Chceme–li určit pravděpodobnost, že se bude náhodná kompozice $\mathbf{X}$ realizovat v $A \subset \mathcal{S}^D$, je třeba vyjádřit $A$ v souřadnicích, tj. $ilr(A)$, a následně spočítat pravděpodobnost z $(D-1)$–rozměrného integrálu, tedy

$$\mathsf{P}(A) = \int\limits_{ilr(A)} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}} (2\pi)^{(D-1)/2}} \exp\left(-\frac{1}{2} (\mathbf{x}^* - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}^* - \boldsymbol{\mu})\right) d\mathbf{x}^*.$$

**Příklad 5** *Uvažujme dvousložkovou kompozici $(X, Y)'$, tedy $D = 2$ a simplex $\mathcal{S}^2$ je pro $k = 1$ ve tvaru*

$$\mathcal{S}^2 = \left\{(x, y) \in \boldsymbol{R}^2, x > 0, y > 0, x + y = 1\right\}.$$

*Nechť množina $A$ je dána jako*

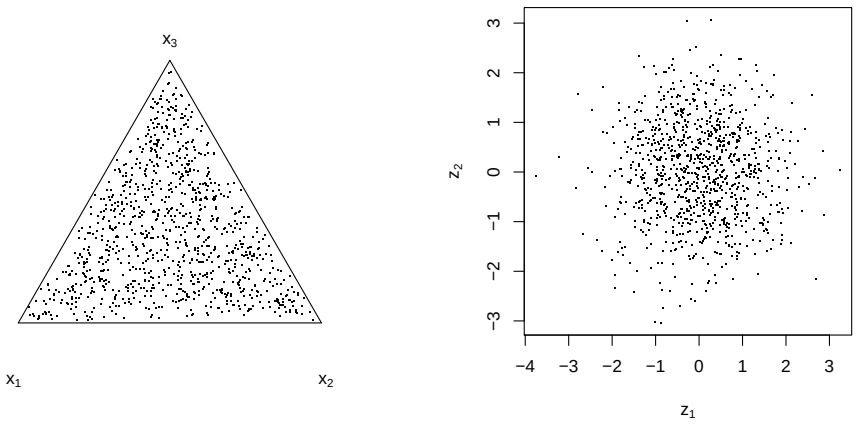
$$A = \left\{(x, y) \in \boldsymbol{R}^2 : x + y = 1;\ 0,25 \leq x \leq 0,75;\ 0,25 \leq y \leq 0,75\right\}$$

*a nechť naším úkolem je určit pravděpodobnost $\mathsf{P}((X, Y)' \in A)$, že se bude kompozice $(X, Y)'$ realizovat v množině $A$ za předpokladu, že se kompozice $(X, Y)'$ řídí normálním rozdělením s parametry $\mu = 1$ a $\sigma^2 = 4$. Určíme nejprve podobu $A$ v souřadnicích. Je zřejmé, že v našem případě stačí transformovat pouze krajní body. Pro výše zvolenou ortonormální bázi na simplexu obdržíme souřadnici $z = \frac{1}{\sqrt{2}} \ln \frac{x}{y}$, a tedy*

$$ilr(A) = \left\{z \in R, \frac{1}{\sqrt{2}} \ln \frac{0,25}{0,75} \leq z \leq \frac{1}{\sqrt{2}} \ln \frac{0,75}{0,25}\right\} =$$

$$= \{z \in R, -0,7768 \leq z \leq 0,7768\}.$$

*Vzhledem k uvedenému rozdělení náhodné veličiny $Z$ je hledaná pravděpodobnost*



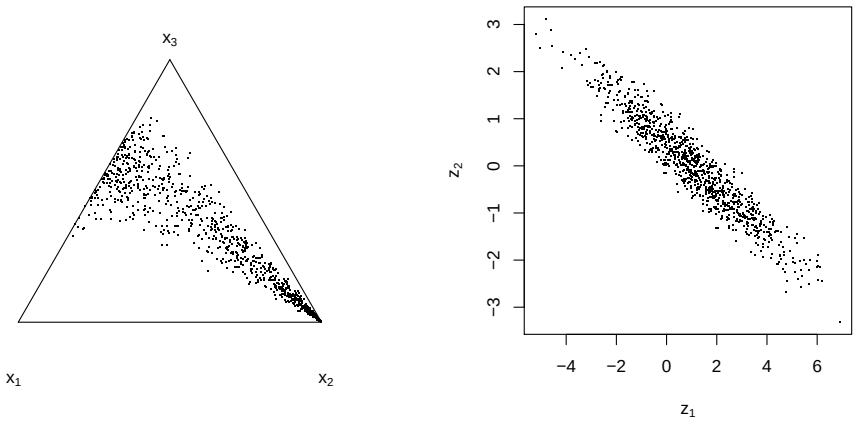

Obrázek 2: Normované normální rozdělení na simplexu, vlevo v ternárním diagramu, vpravo v souřadnicích.

$$\mathsf{P}((X,Y)' \in A) = \mathsf{P}(Z \in ilr(A)) = \int_{-0,7768}^{0,7768} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}} dx = \int_{-0,7768}^{0,7768} \frac{1}{2\sqrt{2\pi}} e^{-\frac{(z-1)^2}{8}} dx$$

$$\Phi\left(\frac{0,7768-1}{2}\right) - \Phi\left(\frac{-0,7768-1}{2}\right) = \Phi\left(-0,1116\right) - \Phi\left(-0,8884\right) =$$

$$= \Phi\left(0,8884\right) - \Phi\left(0,1116\right) = 0,811 - 0,544 = 0,267$$

*Je tedy zřejmé, že hlavní myšlenkou bylo převedení výpočtu pravděpodobnosti do souřadnic, kde již můžeme s výhodou využít známých postupů.*




Obrázek 3: Normální rozdělení na simplexu s parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$, vlevo situace v ternárním diagramu, vpravo v souřadnicích.

# Závěr

Téma mnohorozměrné rozdělení a jeho aplikace jsem si nevybrala jen proto, že je nejznámějším spojitým rozdělením, ale zaujala mě i šíře možností jeho aplikací, kterými bych se ráda v budoucím studiu dále zabývala. Také jsem si chtěla utřídit veškeré pojmy, co se tohoto rozdělení týče a rozšířit znalosti hlouběji, než v samotném kurzu matematické statistiky. Cílem mé bakalářské práce bylo ovšem nejen podrobněji popsat normální rozdělení, aplikaci momentové generující funkce pro jeho jednorozměrný i vícerozměrný případ, ale nakonec seznámit i se základní strukturou kompozičních dat. Jelikož spousta materiálů je v cizím jazyce, představovalo samotné přeložení veškerých materiálů jednu z nejnáročnějších prací, nelehké také bylo naučit se pracovat s dosud pro mě neznámým softwarem TEX. Byla bych ráda, kdyby má bakalářská práce byla přínosem, jak pro mé budoucí studium, tak pro ty, koho příslušné téma zajímá.

# Literatura

[1] Aitchison, J., *The Statistical Analysis of Compositional Data*, Chapman and Hall, London, 1986.

[2] Anděl, J., *Matematická statistika*, SNTL, Praha, 1978.

[3] Egozcue, J.J., Pawlowsky–Glahn, V., Mateu–Figueras, G., Barceló–Vidal, C., *Isometric logratio transformations for compositional data analysis*, Mathematical Geology **35**, 279–300 (2003).

[4] Egozcue, J.J., Pawlowsky–Glahn, V., Mateu–Figueras, G.,*The normal distribution in some constrained sample spaces* [online], dostupné z http://arxiv.org/abs/0802.2643?context=math

[5] Filzmoser, P., Hron, K., *Outlier detection for compositional data using robust methods*, Mathematical Geosciences **40**, 233–248 (2009).

[6] Hogg, R.V., McKean, J.W., Craig, A.T., *Introduction to Mathematical Statistic*, 6. vydání. Prentice Hall, London, 2005.

[7] Jurečková, J., *Úvod do počtu pravděpodobnosti*, Státní pedagogické nakladatelství, Praha, 1972.

[8] Kunderová, P., *Základy pravděpodobnosti a matematické statistiky*, Vydavatelství Univerzity Palackého, Olomouc, 2004.

[9] Lecture notes on compositional data analysis [online], dostupné z: http://hdl.handle.net/10256/297

[citováno 28.5.2007]